Characterization of Audiovisual Dramatic Attitudes

Adela Barbulescu, Rémi Ronfard, Gérard Bailly

▶ **To cite this version:**

Adela Barbulescu, Rémi Ronfard, Gérard Bailly. Characterization of Audiovisual Dramatic Attitudes. Interspeech 2016 - 17th Annual Conference of the International Speech Communication Association, Sep 2016, San Francisco, United States. pp.585-589, ⟨10.21437/Interspeech.2016-75⟩. ⟨hal-01337077⟩

HAL Id: hal-01337077

https://inria.hal.science/hal-01337077v1

Submitted on 24 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Characterization of Audiovisual Dramatic Attitudes

Adela Barbulescu[1]*, Rémi Ronfard*[1]*, Gérard Bailly*[2]

[1]LJK, INRIA & Univ. Grenoble Alpes, Grenoble, France
[2]GIPSA-lab, CNRS & Univ. Grenoble Alpes, Grenoble, France
adela.barbulescu@inria.fr, remi.ronfard@inria.fr, gerard.bailly@gipsa-lab.grenoble-inp.fr

Abstract

In this work we explore the capability of audiovisual parameters (such as voice frequency, rhythm, head motion or facial expressions) to discriminate among different dramatic attitudes. We extract the audiovisual parameters from an acted corpus of attitudes and structure them as frame, syllable, and sentence-level features. Using Linear Discriminant Analysis classifiers, we show that sentence-level features present a higher discriminating rate among the attitudes and are less dependent on the speaker than frame and sylable features. We also compare the classification results with the perceptual evaluation tests, showing that voice frequency is correlated to the perceptual results for all attitudes, while other features, such as head motion, contribute differently, depending both on the attitude and the speaker.
Index Terms: attitude characterization, expressive audiovisual contours, complex emotion

1. Introduction

Attitudes refer to the expression of social affects, which is one of the main goals of speech. The acoustic and visual manifestations of attitudes are linked to conventions and cultural behaviors [1], thus differing from basic emotional expressions, which may be seen as more spontaneous and universal [2].

The study of audiovisual parameters which encode the paralinguistic content of speech plays an essential role in improving the recognition and synthesis of expressive audiovisual speech. To this goal, there has been a great amount of work on the analysis and modeling of features which are found to help in the discrimination between expressive styles. Audiovisual features such as voice quality [3], acoustic prosodic features (F0, rhythm, energy) [4] [5] [6] [7], head motion [8] [9] [10], eye motion [11] and facial expressions [12] [13] [14] [15], have proven to be efficient in discriminating between basic emotions, attitudes or even speaker styles.

While recognition of emotion or psycho-physiological state is largely based on signal-based data mining and deep learning with features collected with a sliding window over multimodal frames [16] [17] [18] [19] , early studies have proposed that speakers use global prosodic patterns to convey an attitude [20][21]. These patterns are supposed to be anchored on the discourse and its linguistic structure, rather than encoded independently on parallel multimodal features. We recently evidenced the relevance of such patterns in facial displays [22].

This work does not focus on proposing an exhaustive list of features for audiovisual emotion recognition, but rather explores the effectiveness of using audiovisual features at different structural levels to discriminate among expressive styles and speakers. We thus compare below the discrimination between attitudes at frame, syllable and sentence-level and with different acoustic and visual features in order to evaluate the importance of the positioning of discriminant audiovisual events within the utterance. To that purpose, we performed a series of Linear Discriminant Analysis (LDA) on an expressive corpus of dramatic attitudes. In line with Iriondo et al [23] who used the results of a subjective test to refine an expressive dataset, we compare our best classification results with perceptual evaluation tests for the set of attitudes which are best discriminated.

Section 2 presents our corpus of attitudes and the extraction and representation of audiovisual features. Section 3 presents the experiments we carried for automatic classification, perceptual evaluation and comparison techniques.

2. Corpus of dramatic attitudes

We use audiovisual data from a corpus of 16 attitudes. The recording and annotation processes are described in [22]. The chosen attitudes represent a subset of the Baron-Cohen's Mind Reading project [24], which proposes a taxonomy of 412 complex emotions. The attitudes were performed by two semi-professional native French actors (one male and one female) under the active supervision of one theater director. In front of a Kinect camera, they were asked to perform 35 utterances (with a technique similar to the *Exercises in Style* by Raymond Queneau [25]) in the following attitudes: declarative (DC), exclamative (EX), question (QS), comforting (CF), tender (TE), seductive (SE), fascinated (FA), jealous (JE), thinking (TH), doubtful (DI), ironic (SH), scandalized (SS), dazed (SD), responsible (RE), confronted (HC) and embarrassed (EM). Given the posed context and the complexity of the expressive categories, we denominate these attitudes as *dramatic*.

2.1. Feature extraction

Voice frequency (F_0). We automatically aligned all utterances with their phonetic transcription and further checked and corrected the automatic estimation of melody by hand using Praat [26]. Therefore, we obtained reliable F_0 contours which we further normalized and then converted to semitones.

Rhythm (Rth). For rhythm we used a duration model [27] [28] where syllable lengthening/shortening is characterized with a unique z-score model applied to log-durations of all constitutive segments of the syllable. We compute a coefficient of lengthening/shortening C corresponding to the deviation of the syllable duration Δ relative to an expected duration Δ':

$$C = \frac{\Delta - \Delta'}{\Delta'} \tag{1}$$

$$\Delta' = (1 - r) \cdot \sum_i \bar{d_{p_i}} + r \cdot D \tag{2}$$

where i is the phoneme index within the syllable, $\bar{d_{p_i}}$ is the average duration of phoneme i, D is the average syllabic duration (=190ms here) and r is a weighting factor for isochronicity (=0.6 here). We note C as the rhythm coefficient which is computed for every syllable in all sentences in the corpus.

Energy (Enr). Energy is extracted at phone-level and computed as mean energy (dB).

Spectrum (Spec). The audio features are extracted using the STRAIGHT vocoder [29]. We use 24 mel-cepstral coefficients, from the 2^{nd} to the 25^{th} (i.e. excluding the energy).

Motion (Head, Gaze). Head and gaze motion are obtained directly from the processing of the Kinect RGBD data by the Faceshift ®software and processed at 30 frames/s.

Facial expressions (Up, Low). Facial expressions are returned by the Faceshift software as blendshape values, to which we apply Non-negative Matrix Factorization (NMF) [30]. We split these into two main groups: *upper-face expressions* (6 components) and *lower-face expressions* (6 components).

2.2. Feature stylization

By *stylization* we mean the extraction of several values at specific locations from the feature trajectories with the main purpose of simplifying the analysis process while maintaining a constant number of characteristics of the original contour for all structural levels whatever the linguistic content. We propose the following stylization methods:

- *audio*: the audio feature contours are stylized by extracting 3 values: at at 20%, 50% and 80% of the vocalic nucleus of each syllable.
- *visual*: the visual feature contours are stylized by extracting contour values at 20%, 50% and 80% of the length of each syllable.
- *rhythm*: the rhythm is represented by one parameter per syllable: the lengthening/shortening coefficient.

3. Experiments

3.1. Discriminant analysis

Discriminant analysis between the 16 attitudes is performed using Fisher classification with 10-fold cross-validation. Speaker-dependent and speaker-independent classification of attitudes were performed at three structural levels for each feature separately and for the concatenation of all audiovisual features:

- *frame-level*: data extracted from each stylization point
- *syllable-level*: concatenation of the frame-level features at each syllable
- *sentence-level*: concatenation of the first and last syllable-level features for a sentence. For sentences composed of one syllable, we perform data duplication to obtain the desired feature dimension (see table 1).

Table 1: Dimension and size for all features.

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
Dimension	1	1	1	24	6	2	6	6	47
Frame sz	1	-	-	1	1	1	1	1	1
Syllable sz	3	1	1	3	3	3	3	3	3
Sentence sz	6	2	2	6	6	6	6	6	6

Results. We observe that the discrimination rate increases as feature granularity increases (see figure 1). Higher scores at sentence-level indicate that order matters: the overall shapes of the features within the sentences have better discrimination power than local feature values. For F_0, head motion and gaze the average F1-score is increased by more than 30% of the scores obtained for the frame- and syllable-level. In the case of spectrum the gain is smaller, showing that it already contains enough discriminant information at frame- and syllable-level.

Individual features generally show a higher score for speaker-dependent classification. It is especially the case for F_0, head motion and face expressions, showing that these gestures manifest different strategies for attitude expression. However, the advantage of structuring features at sentence-level is especially significant in the speaker-independent case. Surprisingly, for the concatenated feature we observe similar scores at sentence-level for all classifiers, while the frame- and syllable-level scores are higher for the speaker-dependent classifiers. This shows there may be a benefit in using speaker-independent models with sentence-level features.

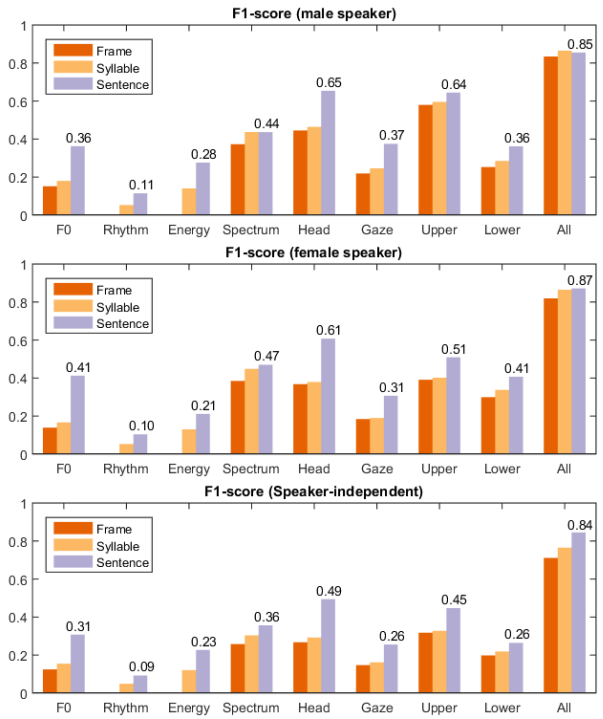

Figure 1: Average F1-scores obtained for all features at frame, syllable and sentence-level, for the male (top), the female speaker (middle) and speaker-independent (bottom). Marked values represent mean F1-scores for sentence-level features.

Table 2 presents the F1-scores for all features at sentence-level. For both speakers, we observe that the highest scores are obtained for head movements, followed by upper-face expressions and voice spectrum. The following features appear in the descending order of the scores: gaze, F_0, lower-face expressions for the male speaker and F_0, lower-face expressions and gaze for the female speaker. Finally, the lowest scoring features are energy and rhythm for both actors. Such observations related to the discrimination capacity of individual features can be used to improve the generation of expressive audiovisual features for discrete attitudes and thus, to expressive audiovisual speech synthesis.

Table 2: F1-scores for the automatic classification. LDA classifiers are trained using sentence-level features over 16 attitudes: for the male speaker (a) and female speaker (b). Values in bold are greater than 0.6.

(a) F1 score for the male speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
Declarative	0.49	0	0.29	**0.61**	**0.82**	0.43	**0.70**	0.08	**0.77**
Exclamative	0.43	0.15	0.10	0.27	**0.70**	0.33	0.58	0.28	**0.72**
Interrogative	0.48	0.36	0.28	0.56	**0.82**	**0.65**	**0.88**	0.29	**0.94**
Comforting	0.08	0.16	0.14	0.20	**0.66**	0.42	0.44	0.36	**0.79**
Tender	0.43	0	0.40	0.47	**0.73**	0.33	0.54	0.56	**0.95**
Seductive	0.30	0.03	0.23	0.47	0.59	0.39	0.55	0.39	**0.87**
Fascinated	0.34	0	0.13	**0.67**	**0.67**	0.41	**0.78**	0.30	**0.94**
Jealous	0.15	0.08	0.19	0.33	0.42	0.19	0.27	0.53	**0.70**
Thinking	0.58	0.40	0.22	0.41	**0.72**	0.41	**0.75**	0.13	**0.89**
Doubtful	0.27	0.16	0.11	0.42	**0.66**	0.31	**0.62**	0.45	**0.88**
Ironic	0.11	0	0.12	0.59	**0.77**	0.44	**0.77**	0.53	**0.89**
Scandalized	0.46	0.09	**0.82**	0.25	**0.66**	0.41	0.59	0.37	**0.85**
Dazed	0.46	0.08	0.31	0.38	0.52	0.16	**0.78**	0.29	**0.76**
Responsible	0.37	0.11	0.41	0.26	**0.60**	0.36	**0.69**	0.39	**0.72**
Confronted	0.45	0	0.10	0.28	**0.69**	0.28	0.44	0.21	**0.80**
Embarrassed	0.37	0.18	**0.62**	**0.90**	0.56	**0.71**	**0.79**	**0.63**	**1**
Mean	0.36	0.11	0.28	0.44	**0.65**	0.37	**0.64**	0.36	**0.85**

(b) F1 score for the female speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
Declarative	0.29	0	0.20	0.55	**0.73**	0.36	**0.80**	0.37	**0.83**
Exclamative	0.31	0	0.14	0.43	**0.68**	0.36	**0.61**	0.19	**0.82**
Interrogative	**0.71**	0.07	0.34	**0.77**	**0.60**	0.07	0.39	0.31	**0.97**
Comforting	0.11	0.10	0.17	**0.61**	0.49	0.34	0.28	0.51	**0.84**
Tender	0.56	0.04	0.31	0.36	0.54	0.26	0.53	**0.69**	**0.94**
Seductive	0.36	0.08	0.24	0.40	0.50	0.16	0.38	0.48	**0.82**
Fascinated	0.36	0	0.12	**0.67**	**0.70**	0.44	**0.61**	0.26	**0.94**
Jealous	0.25	0	0.04	0.41	0.48	0.19	**0.65**	0.18	**0.90**
Thinking	0.31	0.24	0.04	0.37	**0.63**	0.37	0.58	0.27	**0.94**
Doubtful	**0.66**	0	0.15	0.36	0.58	0.37	0.49	0.25	**0.78**
Ironic	0.37	0	0.16	0.22	0.47	0.26	0.36	0.42	**0.75**
Scandalized	**0.61**	0.20	**0.83**	**0.66**	0.47	0.20	0.46	0.45	**0.86**
Dazed	0.36	0.19	0.11	0.31	**0.61**	0.12	0.37	0.30	**0.71**
Responsible	0.31	0.32	0.25	0.39	**0.67**	**0.60**	0.53	0.23	**0.91**
Confronted	**0.68**	0.07	0.19	0.58	**0.73**	0.35	0.59	**0.72**	**0.96**
Embarrassed	0.33	0.24	0.31	0.44	**0.73**	0.58	0.56	**0.94**	**0.97**
Mean	0.41	0.10	0.21	0.47	**0.61**	0.31	0.51	0.41	**0.87**

3.2. Perceptual test

In order to assess the perceptual correlates of these features we carried an attitude recognition test using recorded data from the two speakers. The stimuli were obtained using an animation system, in which the recorded motion is directly mapped to a cartoon-style 3D model of the speaker (see figure 2) and the audio signal is represented by the original voice recordings. The shape deformation of the avatar corresponds exactly to the analyzed visual features, while appearance is represented by cartoon style textures, thus justifying the comparison between the objective and the perceptual test. We focused on a subset of 8 attitudes (Tender, Seductive, Fascinated, Jealous, Thinking, Ironic, Scandalized and Embarrassed) that are best produced and perceived (see [22]).

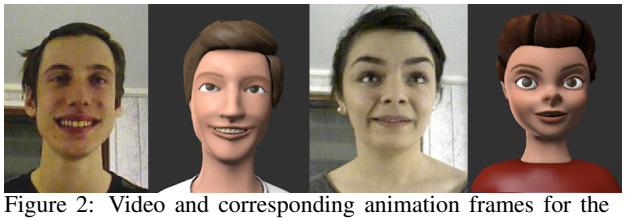

Figure 2: Video and corresponding animation frames for the two speakers performing the attitude Seductive.

The test consists of a set of 32 animations such that no two consecutive performances contain identical attitudes, sentences or speakers. Random sentences are chosen from a subset of 6 sentences such that each attitude appears twice for each speaker. A total of 63 French native speakers participated in this experience. The online test can be found at [1].

Results. The recognition rates for the two speakers are: 55% (male) and 45% (female). The confusion matrices are presented in table 3. The best recognized attitudes are Seductive, Thinking and Scandalized, and lowest is Jealous for both speakers. Generally, the male speaker was better recognized, this observation being supported by the results obtained in the automatic classification. Notable observations: for the male speaker, Jealous and Ironic are interchangeably confused, while for the female speaker, Jealous is confused with Embarrassed.

[1] http://www.gipsa-lab.fr/∼adela.barbulescu/test_recognition/

Table 3: Confusion matrices obtained for the perception test with performances of the male speaker (top) and female speaker (bottom). Values in bold are greater than 30%.

(%)	TE	SE	FA	JE	TH	IR	SS	EM
TE	**47.7**	8.5	24.0	0.8	6.2	10.0	0.0	3.1
SE	9.2	**52.3**	4.7	7.0	6.9	15.4	1.5	3.1
FA	10.0	3.8	**71.3**	0.0	8.5	0.0	1.5	4.6
JE	2.3	2.3	0.0	**34.9**	6.2	**46.2**	6.2	1.5
TH	0.8	1.5	3.1	3.1	**74.6**	9.2	0.8	6.9
IR	1.5	10.8	0.8	18.6	8.5	**51.5**	3.8	4.6
SS	0.8	0.0	18.6	4.7	0.8	3.1	**68.5**	3.8
EM	20.0	10.8	13.2	0.8	10.8	3.8	0.0	**40.8**

(%)	TE	SE	FA	JE	TH	IR	SS	EM
TE	**37.7**	9.2	**30.8**	1.5	4.7	4.6	5.4	6.2
SE	18.5	**49.2**	4.6	3.1	10.9	5.4	1.5	6.9
FA	13.8	4.6	**30.8**	0.8	**31.8**	0.8	3.1	14.6
JE	9.2	8.5	2.3	17.7	10.9	13.8	6.9	**30.8**
TH	2.3	2.3	1.5	7.7	**50.4**	12.3	3.1	20.0
IR	4.6	2.3	1.5	16.2	14.7	**35.4**	11.5	13.8
SS	0.8	0.0	8.5	2.3	0.8	0.8	**84.6**	2.3
EM	9.2	16.2	3.1	1.5	3.1	3.8	0.0	**63.1**

3.3. Comparison between objective and perceptual scores

In order to compare the discrimination scores obtained by automatic classification and the perceptual test results, we trained separate LDA classifiers for the two speakers with the 8 attitudes used in the perceptual tests. Data was partitioned into training and testing such that the testing sentences coincide with

Table 4: F1-scores and correlation coefficients for the automatic classification and perceptual tests. LDA classifiers are trained using sentence-level features over 8 attitudes and tested on sentences used in perceptual test (Perc): for the male (a) and female speaker (b). Values in bold are greater than 0.6.

(a) F1-score for the male speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All	Perc
TE	0.33	0.35	**0.76**	0.54	**0.91**	0.23	0.50	**0.83**	**0.79**	0.49
SE	**0.63**	0.12	**0.62**	0.50	0.56	0.36	**0.70**	**0.71**	**0.71**	0.54
FA	0.56	0.02	0.20	**0.71**	**0.87**	**0.89**	**0.94**	0.58	**0.99**	**0.61**
JE	0.10	0.42	0.27	0.46	0.47	0.36	0.32	**0.79**	**0.74**	0.40
TH	**0.67**	0.22	0.05	0.20	**0.86**	0.46	**0.89**	0.33	**0.91**	**0.67**
IR	0	0	**0.70**	0.58	**0.80**	**0.89**	**0.86**	**0.71**	**0.97**	0.43
SS	**0.66**	0.12	**0.98**	0.59	**0.95**	**0.64**	**0.95**	**0.78**	**0.98**	**0.76**
EM	**0.61**	0.30	**0.69**	**0.74**	0.55	0.60	**0.93**	**0.84**	**0.92**	0.50
Mean	0.45	0.19	0.54	0.54	**0.75**	0.55	**0.76**	**0.70**	**0.88**	0.55

(b) F1-score for the female speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All	Perc
TE	**0.67**	0	0.56	0.43	**0.71**	0.26	**0.93**	**0.96**	**0.80**	0.38
SE	0.40	0.30	0.53	0.32	**0.71**	0.08	0.38	**0.87**	**0.67**	0.51
FA	0.34	0.02	0.16	0.36	**0.73**	**0.72**	**0.91**	0.41	**0.83**	0.33
JE	0.07	0.09	0.46	0.22	0.32	0.49	0.48	0.49	**0.82**	0.23
TH	**0.76**	0.26	0.02	0.27	0.38	0.55	0.55	0.24	**0.69**	0.44
IR	0.35	0	0.26	0.56	0.29	0.24	0.19	**0.62**	**0.80**	0.40
SS	**0.98**	0.30	**0.99**	0.55	0.59	0.44	**0.74**	**0.62**	**0.95**	**0.79**
EM	0.39	0.37	**0.76**	0.16	**0.76**	**0.75**	**0.84**	**0.98**	**0.96**	0.49
Mean	0.50	0.17	0.47	0.36	0.56	0.44	**0.63**	**0.65**	**0.82**	0.45

Table 5: Correlation coefficients for the automatic classification and perceptual tests for the male (a) and female speaker (b). Values in bold are associated with p-values < 0.1.

(a) Correlations for male speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
r	**0.78**	-0.3	0.04	-0.1	**0.62**	0.16	**0.62**	-0.4	0.50
p	**0.02**	0.38	0.91	0.68	**0.09**	0.69	**0.09**	0.26	0.20

(b) Correlations for female speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
r	**0.78**	**0.64**	**0.65**	0.40	0.27	-0.1	0.10	0.20	0.35
p	**0.02**	**0.08**	**0.07**	0.31	0.50	0.79	0.80	0.62	0.38

the ones used in the perceptual test. Table 4 shows the F1-scores obtained by automatic classification for all features at sentence-level and the scores obtained for the perceptual tests.

As expected, the perceptual test scores are generally lower than the automatic classification scores. This phenomena has multiple causes, such as: the acting strategy may not correspond to the subjects' view of that attitude, the decrease of expressiveness introduced by animation etc. For the objective and perceptual tests, the highest recognized attitude is Scandalized, while the least recognized is Jealous. As in the case of automatic classification for 16 attitudes, the male speaker generally obtained higher scores and this is reflected also in the perceptual test.

We computed the Pearson correlation between the F1-scores obtained for the tests (see table 5). We set the significance level at 0.1 and obtained strong correlations for F_0 for both speakers. Moderate correlations are also obtained for the female speaker for rhythm, and for the male speaker for head motion and upper-face expressions. The male speaker seems to make a better use of head motion to encode attitudes and this is perceived and effectively used by viewers. However, the use of recorded voices with synthetic visual input may explain the dominance of F_0 correlations in the perceptual scores. Figure 3 illustrates attitude discriminant stylized contours.

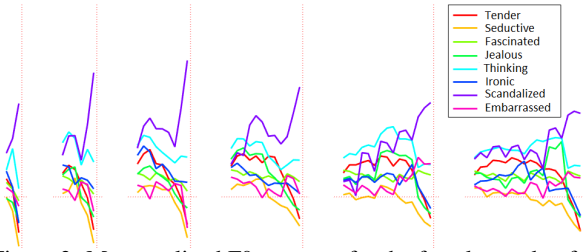

Figure 3: Mean stylized F0 contours for the female speaker for the 8 attitudes as a function of number of syllables of the utterance (here 1 to 6 syllables). This evidences that Scandalized and Thinking patterns strongly stand out from the others.

4. Conclusion

We analyzed audiovisual speech utterances with similar content (sentences) in different styles (dramatic attitudes). In a series of experiments, we found that LDA classifiers trained on individual features at sentence-level outperform classifiers trained on frame- or syllable-level features, thus showing a benefit in introducing the temporal dimension for audiovisual features at the level of the sentence for the recognition and synthesis of attitudes. Among these features, head motion and facial expressions show higher discrimination scores than F_0.

Moreover, we observed that the usage of sentence-level features brings the highest gains in score for speaker-independent classifiers. This finding encourages future work in the direction of building speaker-independent models of attitudes with sentence-level features. As previous studies show that preparatory movements help in the discrimination of expressive styles [8], the analysis could also include motion data extracted from pre- and post-phonatory silences.

By analyzing individual features at sentence-level, we found that F_0 is strongly correlated with perceptual judgements, while for other features, correlations are found only for one speaker. Such an example is the effective usage of head motion by the male speaker. Future work will focus on analyzing the individual contributions of visual features in the perception of attitudes, similar to the approach presented in [31]. For this, we are creating synthetic stimuli by individually controlling visual features at sentence level using our 3-D animation platform.

5. Acknowledgements

This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the European Research Council advanced grant EXPRESSIVE (ERC-2011-ADG 20110209).

6. References

[1] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[2] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[3] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, "Discriminating expressive speech styles by voice quality parameterization," in *Proc. of ICPhS*, 2007.

[4] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in french: data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[5] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, Jan. 2003.

[6] I. Iriondo, S. Planet, J.-C. Socoró, and F. Alías, "Objective and subjective evaluation of an expressive speech corpus," in *Advances in Nonlinear Speech Processing*. Springer, 2007, pp. 86–94.

[7] H. Mixdorff, A. Hönemann, and A. Rilliard, "Acoustic-prosodic analysis of attitudinal expressions in german," *Proceedings of Interspeech 2015*, 2015.

[8] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, ser. FGR '02, Washington, DC, USA, 2002, pp. 396–.

[9] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1075–1086, 2007.

[10] A. Barbulescu, R. Ronfard, G. Bailly, G. Gagneré, and H. Cakmak, "Beyond basic emotions: expressive virtual actors with social attitudes," in *Proceedings of the Seventh International Conference on Motion in Games*. ACM, 2014, pp. 39–47.

[11] K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell, "Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems," in *Eurographics 2014 - State of the Art Reports*, France, Apr. 2014, pp. 69–91.

[12] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[13] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[14] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '13. USA: ACM, 2013, pp. 25–35.

[15] C. Davis, J. Kim, V. Aubanel, G. Zelic, and Y. Mahajan, "The stability of mouth movements for multiple talkers over multiple sessions," *Proceedings of the 2015 FAAVSP*.

[16] R. E. Kaliouby and P. Robinson, "Mind reading machines: automated inference of cognitive mental states from video," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 1, 2004, pp. 682–688.

[17] E. Chuang and C. Bregler, "Mood swings: Expressive speech animation," *ACM Trans. Graph.*, vol. 24, no. 2, pp. 331–347, Apr. 2005.

[18] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[19] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[20] I. Fónagy, E. Bérard, and J. Fónagy, "Clichés mélodiques," *Folia linguistica*, vol. 17, no. 1-4, pp. 153–186, 1983.

[21] D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[22] A. Barbulescu, G. Bailly, R. Ronfard, and M. Pouget, "Audiovisual generation of social attitudes from neutral stimuli," in *Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[23] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

[24] S. Baron-Cohen, *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers, 2003.

[25] R. Queneau, *Exercises in style*. New Directions Publishing, 2013.

[26] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[27] W. N. Campbell, "Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[28] G. Bailly and B. Holm, "Sfc: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[29] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[30] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source code for biology and medicine*, vol. 8, no. 1, p. 1, 2013.

[31] A. Adams, M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Decoupling facial expressions and head motions in complex emotions," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 274–280.